Day 9: Topic Models

ME414: Introduction to Data Science and Big Data Analytics

LSE Methods Summer Programme

24 August 2017

Day 9 Outline

Probabilistic topic models

Latent Dirichlet allocation (LDA)

Beyond Latent Dirichlet Allocation

Correlated and Dynamic Topic Models

Supervised Topic Models
- Relational topic models
- Ideal point topic models

Extending LDA

Bayesian Nonparametric Models

Using and Checking Topic Models

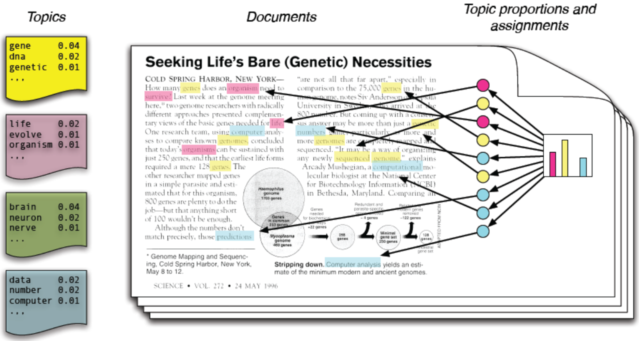
Probabilistic topic models

Probabilistic topic models

- Topic modeling allows us to automatically organize, understand, and summarize large archives of text data.
- Uncover hidden themes.
- Annotate the documents according to themes.
- Organize the collection using annotations.

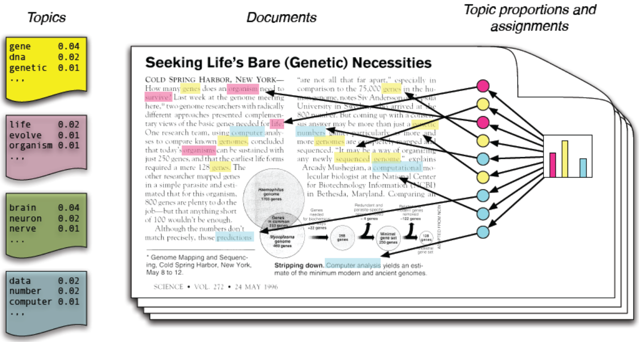
Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation (LDA)



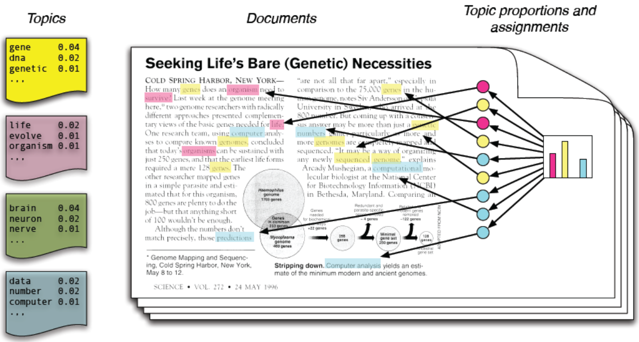
- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

Latent Dirichlet allocation (LDA)



- In reality, we only observe the documents
- The other structure are hidden variables

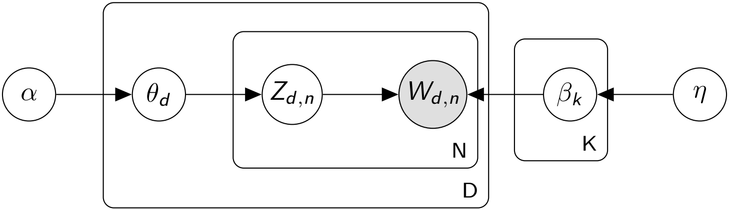
Latent Dirichlet allocation (LDA)



- Our goal is to infer the hidden variables
- I.e., compute their distribution conditioned on the documents

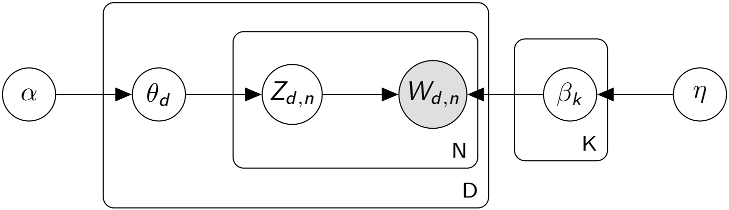
$$p(topics, proportions, assignments | documents)$$

LDA as a graphical model



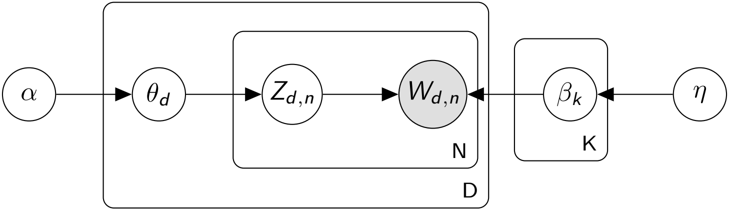
- Encodes assumptions
- Defines a factorization of the joint distribution
- Connects to algorithms for computing with data

LDA as a graphical model



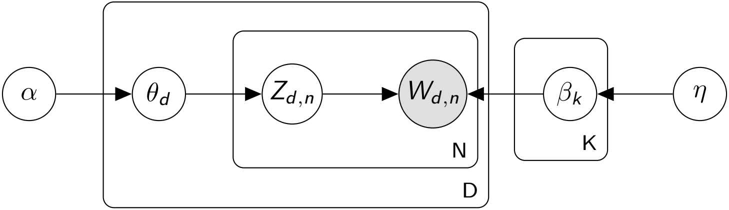
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

LDA as a graphical model



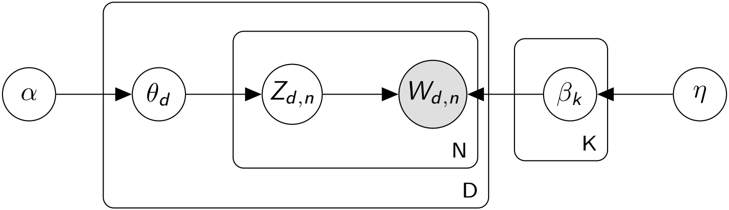
- α proportions parameter
- θ_d per-document topic proportions
- $Z_{d,n}$ per-word topic assignment
- $W_{d,n}$ observed word
- β_k topics
- η topic parameter

LDA as a graphical model



$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i | \eta) \right) \left(\prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:k}, z_{d,n}) \right)$$

LDA as a graphical model



- This joint defines a posterior, $p(\theta, z, \beta | w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.

The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

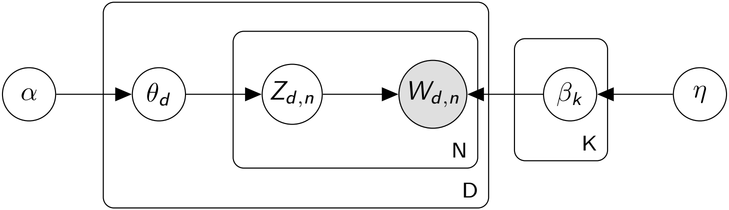
$$p(\theta|\vec{\alpha}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- It is conjugate to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ.
- The topic proportions are a K dimensional Dirichlet. The topics are a V dimensional Dirichlet.

Why does LDA "work"?

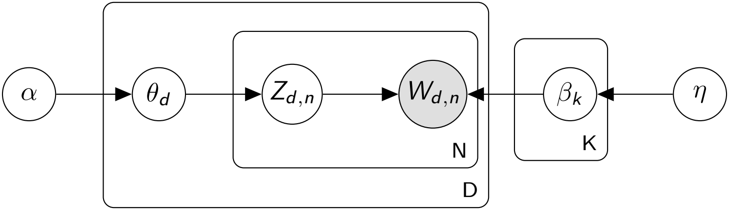
- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible.
  2. For each topic, assign high probability to as few terms as possible.
- These goals are at odds.
  - Putting a document in a single topic makes (2) hard: All of its words must have probability under that topic.
  - Putting very few words in each topic makes (1) hard: To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

LDA summary



- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a posterior inference problem.
- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.
- Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999). It is a mixed-membership model (Erosheva, 2004). It relates to PCA and matrix factorization (Jakulin and Buntine, 2002). It was independently invented for genetics (Pritchard et al., 2000).

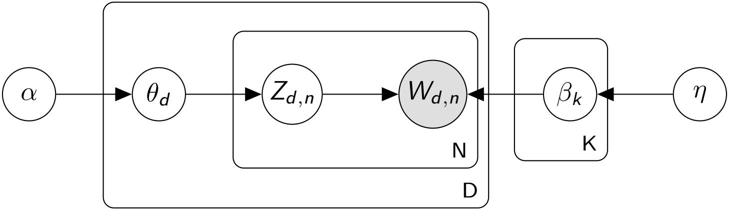
LDA summary



- LDA is a simple building block that enables many applications.
- It is popular because organizing and finding patterns in data has become important in the sciences, humanities, industry, and culture.
- Further, algorithmic improvements let us fit models to massive data.

Beyond Latent Dirichlet Allocation

LDA summary



- LDA is a simple topic model.
- It can be used to find topics that describe a corpus.
- Each document exhibits multiple topics.
- There are several ways to extend this model.

Extending LDA

- LDA can be embedded in more complicate models, embodying further intuitions about the structure of the texts.
- E.g., it can be used in models that account for syntax, authorship, word sense, dynamics, correlation, hierarchies, and other structure.
- The data generating distribution can be changed. We can apply mixed-membership assumptions to many kinds of data.
- E.g., we can build models of images, social networks, music, purchase histories, computer code, genetic data, and other types.
- The posterior can be used in creative ways.
- E.g., we can use inferences in information retrieval, recommendation, similarity, visualization, summarization, and other applications.

Extending LDA

- These different kinds of extensions can be combined.
- To give a sense of how LDA can be extended, we'll look at several examples of major extensions.
- We will discuss
  - Correlated topic models
  - Dynamic topic models & measuring scholarly impact
  - Supervised topic models
  - Relational topic models
  - Ideal point topic models
  - Collaborative topic models

Correlated and Dynamic Topic Models

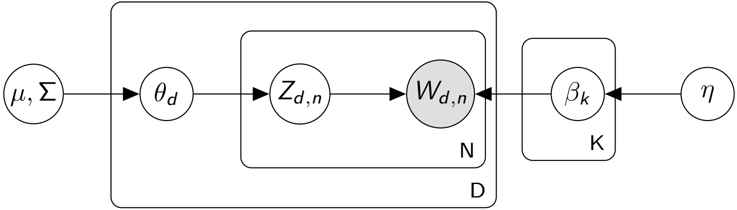
Correlated topic models

- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about fossil fuels is more likely to also be about geology than about genetics.
- The logistic normal is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$X \sim N_k(\mu, \Sigma)$$

$$\theta_i \propto exp\{x_i\}.$$

Correlated topic models



where the first node is logistic normal prior.

- Draw topic proportions from a logistic normal.
- This allows topic occurrences to exhibit correlation.
- Provides a "map" of topics and how they are related
- Provides a better fit to text data, but computation is more complex

Dynamic topic models

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time.
- Dynamic topic models let the topics drift in a sequence.

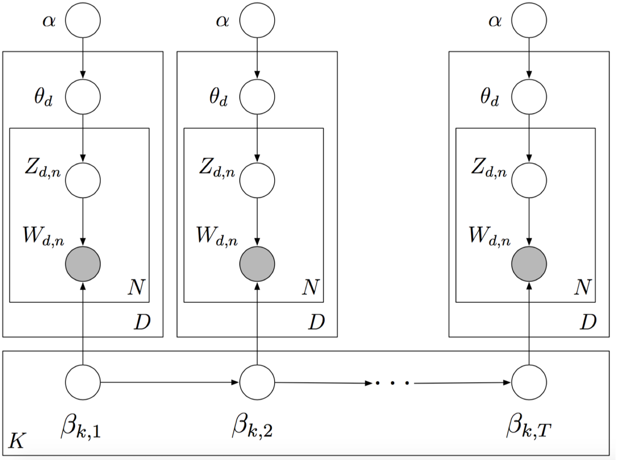


Plate (K) is topics drift through time.

Dynamic topic models

$$\beta_{k,1} \longrightarrow \beta_{k,2} \longrightarrow \cdots \longrightarrow \beta_{k,T}$$

- Use a logistic normal distribution to model topics evolving over time.
- Embed it in a state-space model on the log of the topic distribution

$$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, I\sigma^2)$$
$$p(w|\beta_{t,k}) \propto exp\{\beta_{t,k}\}$$

- As for CTMs, this makes computation more complex. But it lets us make inferences about sequences of documents.

Dynamic topic models

- Time-corrected similarity shows a new way of using the posterior.
- Consider the expected Hellinger distance between the topic proportions of two documents,

$$d_{ij} = E\left[\sum_{k=1}^{K}(\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 | \mathbf{w}_i, \mathbf{w}_j\right]$$

- Uses the latent structure to define similarity
- Time has been factored out because the topics associated to the components are different from year to year.
- Similarity based only on topic proportions

Summary: Correlated and dynamic topic models

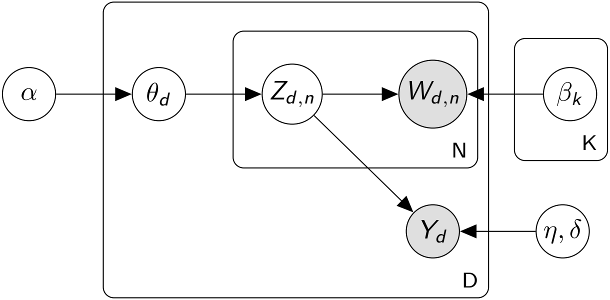
- The Dirichlet assumption on topics and topic proportions makes strong conditional independence assumptions about the data.
- The correlated topic model uses a logistic normal on the topic proportions to find patterns in how topics tend to co-occur.
- The dynamic topic model uses a logistic normal in a linear dynamic model to capture how topics change over time.
- What's the catch? These models are harder to compute.

Supervised Topic Models

Supervised LDA

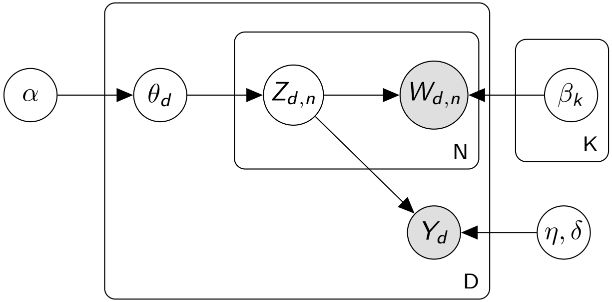
- LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with response variables.
  - User reviews paired with a number of stars
  - Web pages paired with a number of "likes"
  - Documents paired with links to other documents
  - Images paired with a category
- Supervised LDA are topic models of documents and responses. They are fit to find topics predictive of the response.

Supervised LDA



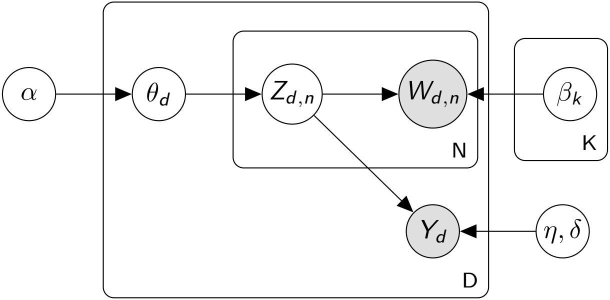
- Y_d is document response
- η, δ regression parameters

Supervised LDA



1. Draw topic proportions $\theta | \alpha \sim Dir(\alpha)$
2. For each word
   - Draw topic assignment $z_n | \theta \sim Mult(\theta)$.
   - Draw word $w_n | z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$
3. Draw response variable $y | z_{1:N}, \eta, \sigma^2 \sim N(\eta^T \bar{z}, \sigma^2)$ where $\bar{z} = (1/N) \sum_{n=1}^{N} z_n$.

Supervised LDA

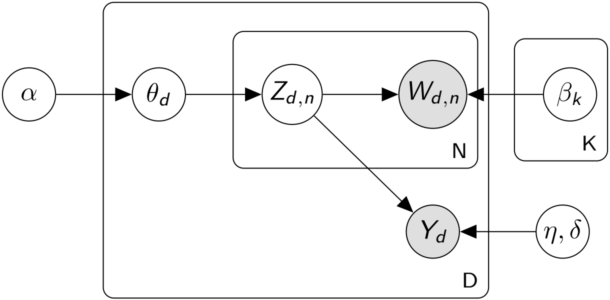


- Fit sLDA parameters to documents and responses. This gives: topics $\beta_{1:K}$ and coefficients $\eta_{1:K}$.
- Given a new document, predict its response using the expected value:

$$E\left[Y|w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2\right] = \eta^T E[\bar{Z}|w_{1:N}]$$

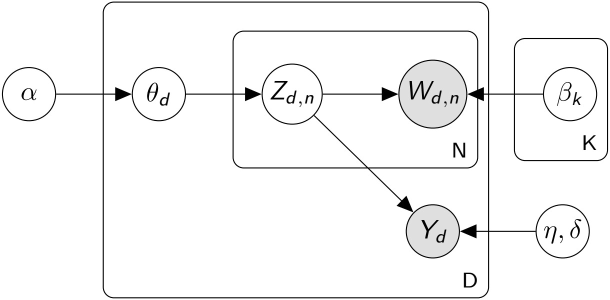
- This blends generative and discriminative modeling.

Supervised LDA



- sLDA enables model-based regression where the predictor is a document.
- It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).
- sLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.

Supervised LDA



- sLDA has been extended to generalized linear models, e.g., for image classification and other non-continuous responses.
- We will discuss two extensions of sLDA
  - Relational topic models: Models of networks and text
  - Ideal point topic models: Models of legislative voting behavior

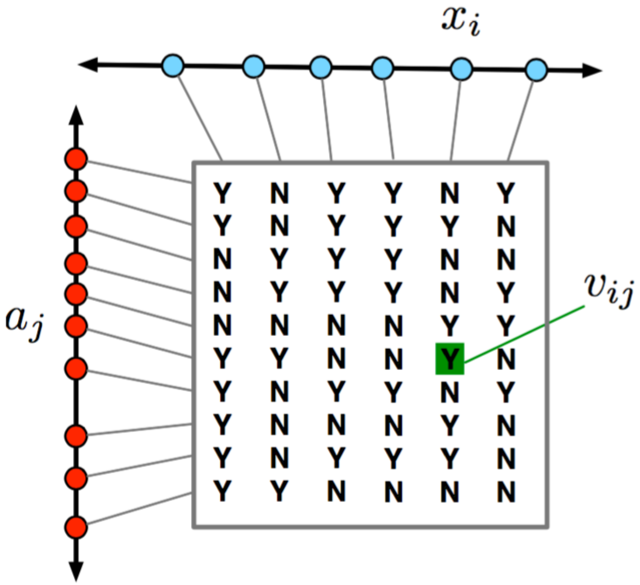
Relational topic models

- Many data sets contain connected observations.
- For example:
  - Citation networks of documents
  - Hyperlinked networks of web-pages.
  - Friend-connected social network profiles

Relational topic models

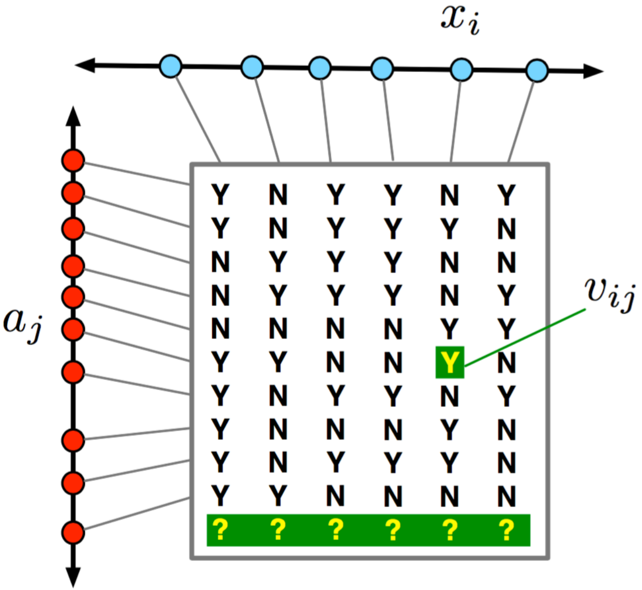
- Research has focused on finding communities and patterns in the link-structure of these networks. But this ignores content.
- sLDA was adapted to pairwise response variables. This leads to a model of content and connection.
- Relational topic models find related hidden structure in both types of data.

Ideal point topic models



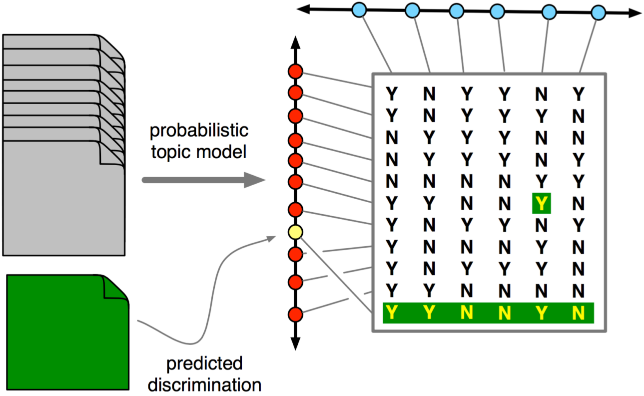
$$p(v_{ij}) = f(d(x_i, a_j))$$

Ideal point topic models



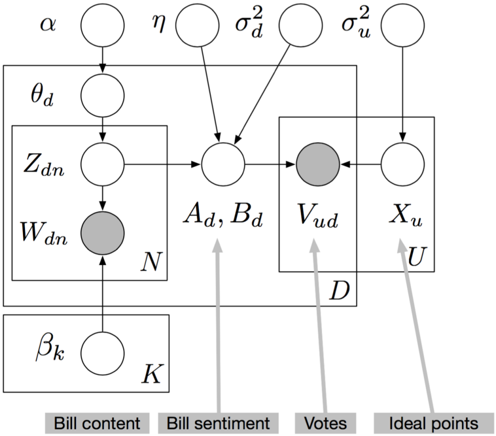
$$p(v_{ij}) = f(d(x_i, a_j))$$

Ideal point topic models



- Use supervised LDA to predict bill discrimination from bill text.
- But this is a latent response.

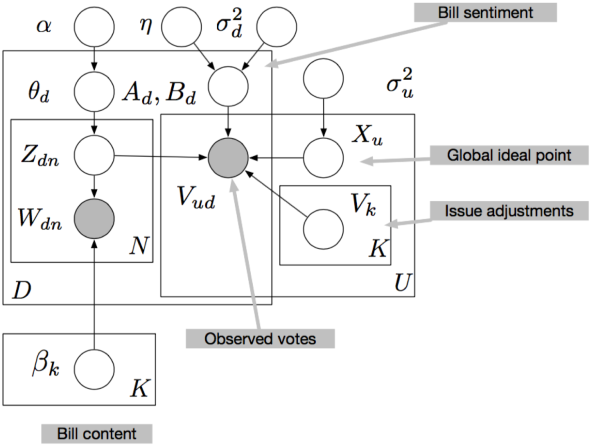
Ideal point topic models



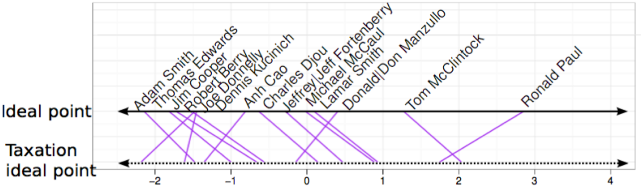
Issue-adjusted ideal points

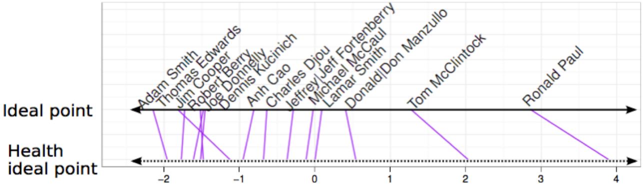
- Ideal point model uses topics to predict votes from new bills.
- Alternatively, we can use the text to characterize how legislators diverge from their usual ideal points.
- For example: A senator might be left wing, but vote conservatively when it comes to economic matters.

Issue-adjusted ideal points



Issue-adjusted ideal points





Summary: Supervised topic models

- Many documents are associated with response variables.
- Supervised LDA embeds LDA in a generalized linear model that is conditioned on the latent topic assignments.
- Relational topic models use sLDA assumptions with pair-wise responses to model networks of documents.
- Ideal point topic models demonstrates how the response variables can themselves be latent variables. In this case, they are used downstream in a model of legislative behavior.
- (sLDA, the RTM, and others are implemented in the R package "lda.")

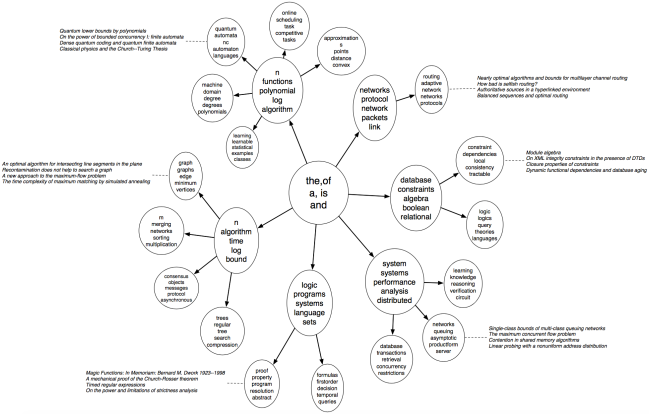
Extending LDA

- Syntactic topic models
- Topic models on images
- Topic models on social network data
- Topic models on music data
- Topic models for recommendation systems
- Spike and slab priors
- Models of word contagion
- N-gram topic models

Bayesian Nonparametric Models

- Topic models assume that the number of topics is fixed.
- It is a type of regularization parameter. It can be determined by cross validation and other model selection techniques.
- Bayesian nonparametric methods skirt model selection:
  - The data determine the number of topics during inference.
  - Future data can exhibit new topics.
- (This is a field unto itself, but has found wide application in topic modeling.)

Hierarchical topic model (Blei et al. 2010)



Summary: Bayesian nonparametrics

- Bayesian nonparametric modeling is a growing field (Hjort et al., 2011).
- BNP methods can define priors over latent combinatorial structures.
- In the posterior, the documents determine the particular form of the structure that is best for the corpus at hand.
- Recent innovations:
  - Improved inference (Blei and Jordan, 2006, Wang et al. 2011)
  - BNP models for language (Teh, 2006; Goldwater et al., 2011)
  - Dependent models, such as time series models (MacEachern 1999, Dunson 2010, Blei and Frazier 2011)
  - Predictive models (Hannah et al. 2011)
  - Factorization models (Griffiths and Ghahramani, 2011)

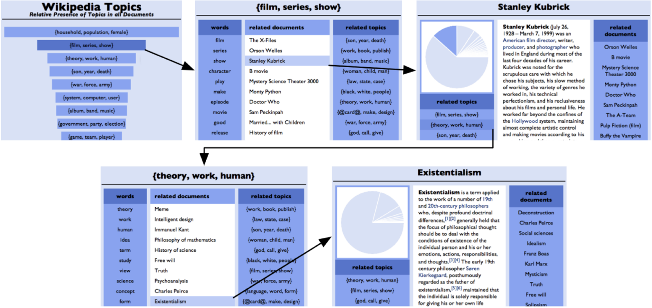
Using and Checking Topic Models

- We have collected data, selected a model, and inferred the posterior.
- How do we use the topic model?
- Using a model means doing something with the posterior inference.
- E.g., visualization, prediction, assessing document similarity, using the representation in a downstream task (like IR).

Using and Checking Topic Models

- Questions we ask when evaluating a model:
  - Does my model work? Is it better than another model?
  - Which topic model should I choose? Should I make a new one?
- These questions are tied up in the application at hand.
- Sometimes evaluation is straightforward, especially in prediction tasks.

Using and Checking Topic Models



- But a promise of topic models is that they give good exploratory tools. Evaluation is complicated, e.g., is this a good navigator of my collection?
- And this leads to more questions:
  - How do I interpret a topic model?
  - What quantities help me understand what it says about the data?

Using and Checking Topic Models

- How to interpret and evaluate topic models is an active area of research.
  - Visualizing topic models
  - Naming topics
  - Matching topic models to human judgements
  - Matching topic models to external ontologies
  - Computing held out likelihoods in different ways
- There are two main components:
  - Predictive scores for evaluating topic models
  - Posterior predictive checks for topic modeling

The predictive score

- Assess how well a model can predict future data.
- In text, a natural setting is one where we observe part of a new document and want to predict the remainder.
- The predictive distribution is a distribution conditioned on the corpus and the partial document,

$$p(w|\mathcal{D}, \mathbf{w}_{obs}) = E_q[\theta|\mathbf{w}_{obs}]^T E_q[\beta_{\cdot,w}|\mathcal{D}].$$

The predictive score

- The predictive score evaluates the remainder of the document independently under this distribution.

$$s = \sum_{w \in \mathbf{w}_{held\ \ out}} log\ \ p(w|\mathcal{D}, \mathbf{w}_{obs})$$

- In the predictive distribution, q is any approximate posterior. This puts various models and inference procedures on the same scale.
- (In contrast, perplexity of entire held out documents requires different approximations for each inference method.)

Posterior predictive checks

- The predictive score and other model selection criteria are good for choosing among several models.
- But they don't help with the model building process; they don't tell us how a model is misfit. (E.g. should I go from LDA to a DTM or LDA to a CTM?)
- Further, prediction is not always important in exploratory or descriptive tasks. We may want models that capture other aspects of the data.
- Posterior predictive checks are a technique from Bayesian statistics that help with these issues.

Posterior predictive checks

- Three stages to model building: estimation, criticism, and revision.
- In criticism, the model "confronts" our data.
- Suppose we observe a data set $\mathbf{y}$. The predictive distribution is the distribution of data *if the model is true*:

$$p(y|M) = \int_{\theta} p(y|\theta)p(\theta)$$

- Locating $\mathbf{y}$ in the predictive distribution indicates if we can "trust" the model.
- Or, locating a discrepancy function $g(\mathbf{y})$ in its predictive distribution indicates if what is important to us is captured in the model.

Posterior predictive checks

- Rubin (1984) located the data $\mathbf{y}$ in the posterior $p(y|\mathbf{y}, M)$.
- Gelman, Meng, Stern (1996) expanded this idea to "realized discrepancies" that include hidden variables $g(\mathbf{y}, \mathbf{z})$.
- We might make modeling decisions based on a variety of simplifying considerations (e.g., algorithmic). But we can design the realized discrepancy function to capture what we really care about.
- Further, realized discrepancies let us consider which parts of the model fit well and which parts don't. This is apt in exploratory tasks.

Posterior predictive checks in topic models

- Consider a decomposition of a corpus into topics, i.e., $\{w_{d,n}, z_{d,n}\}$. Note that $z_{d,n}$ is a latent variable.
- For all the observations assigned to a topic, consider the variable $\{w_{d,n}, d\}$. This is the observed word and the document it appeared in.
- One measure of how well a topic model fits the LDA assumptions is to look at the per-topic mutual information between w and d.
- If the words from the topic are independently generated then we expect lower mutual information.
- What is "low"? To answer that, we can shuffle the words and recompute. This gives values of the MI when the words are independent.

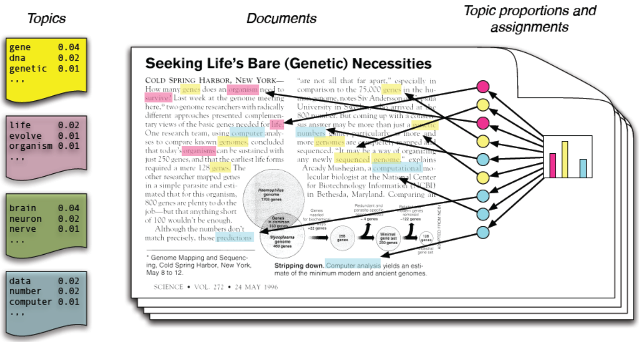
Posterior predictive checks in topic models

- This realized discrepancy measures model fitness.
- Can use it to measure model fitness per topic.
- Helps us explore parts of the model that fit well.

Summary

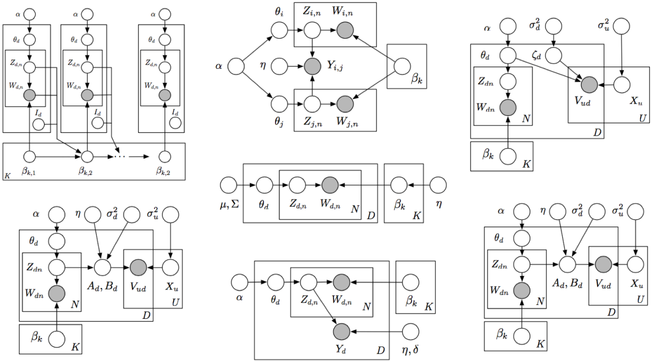
- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- How do I evaluate and check a topic model?
- What are some unanswered questions in this field?
- How can I learn more?

Summary



- LDA assumes that there are K topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.

Summary



Topic models can be adapted to many settings

- relax assumptions
- combine models
- model more complex data

Summary

- Posterior inference is the central computational problem.
- Stochastic variational inference is a scalable algorithm.
- We can handle nonconjugacy with Laplace inference.
- (Note: There are many types of inference and it merits a separate discussion.)

Implementations of LDA

Incomplete list:

- LDA-C – A C implementation of LDA
- HDP – A C implementation of the HDP ("infinite LDA")
- Online LDA – A python package for LDA on massive data
- LDA in R – Package in R for many topic models
- LingPipe – Java toolkit for NLP and computational linguistics
- Mallet – Java toolkit for statistical NLP
- TMVE – A python package to build browsers from topic models